Generalized Method of Moments and Empirical Likelihood

Guido W. IMBENS

Department of Economics and Department of Agricultural and Resource Economics, University of California Berkeley, 649 Evans Hall #3880, Berkeley, CA 94720 (*imbens@econ.berkeley.edu*) and National Bureau of Economic Research

Generalized method of moments (GMM) estimation has become an important unifying framework for inference in econometrics in the last 20 years. It can be thought of as encompassing almost all of the common estimation methods, such as maximum likelihood, ordinary least squares, instrumental variables, and two-stage least squares, and nowadays is an important part of all advanced econometrics textbooks. The GMM approach links nicely to economic theory where orthogonality conditions that can serve as such moment functions often arise from optimizing behavior of agents. Much work has been done on these methods since the seminal article by Hansen, and much remains in progress. This article discusses some of the developments since Hansen's original work. In particular, it focuses on some of the recent work on empirical likelihood–type estimators, which circumvent the need for a first step in which the optimal weight matrix is estimated and have attractive information theoretic interpretations.

KEY WORDS: Empirical likelihood; Generalized method of moments; Moment conditions.

1. INTRODUCTION

Generalized method of moments (GMM) estimation has become an important unifying framework for inference in econometrics in the last 20 years. It can be thought of as encompassing almost all of the common estimation methods such as maximum likelihood, ordinary least squares, instrumental variables, and two-stage least squares, and nowadays is an important part of all advanced econometrics textbooks (Gallant 1987; Davidson and McKinnon 1993; Hamilton 1994; Hayashi 1999; Mittelhammer, Judge, and Miller 2000; Ruud 2000; Wooldridge 2002). Its formalization by Hansen (1982) centers on the presence of known functions, labeled "moment functions," of observable random variables and unknown parameters that have expectation zero when evaluated at the true parameter values. The method generalizes the "standard" method of moments where expectations of known functions of observable random variables are equal to known functions of the unknown parameters. The "standard" method of moments can thus be thought of as a special case of the general method with the unknown parameters and observed random variables entering additively separable. The GMM approach links nicely to economic theory where orthogonality conditions that can serve as such moment functions often arise from optimizing behavior of agents. For example, if agents make rational predictions with squared error loss, then their prediction errors should be orthogonal to elements of the information set. In the GMM framework, the unknown parameters are estimated by setting the sample averages of these moment functions, the "estimating equations," as close to zero as possible.

The framework is sufficiently general to deal with the case where the number of moment functions is equal to the number of unknown parameters, the so-called "just-identified case," as well as the case in which the number of moments exceeds the number of parameters to be estimated, the "overidentified case." The latter has special importance in economics, where the moment functions often come from the orthogonality of potentially many elements of the information set and prediction errors. In the just-identified case, it is typically possible to estimate the parameter by setting the sample average of the moments exactly equal to zero. In the overidentified case, this is not feasible. The solution proposed by Hansen (1982) for this case, following similar approaches in linear models, such as two- and three-stage least squares, is to set a linear combination of the sample average of the moment functions equal to zero, with the dimension of the linear combination equal to the number of unknown parameters. The optimal linear combination of the moments depends on the unknown parameters; Hansen suggested using initial, possibly inefficient, estimates to estimate this optimal linear combination. Chamberlain (1987) showed that this class of estimators achieves the semiparametric efficient bound given the set of moment restrictions. The Chamberlain article is not only important for its substantive efficiency result, but also as a precursor to the subsequent empirical likelihood literature by the methods used: Chamberlain used a discrete approximation to the joint distribution of all of the variables to show that the information matrix–based variance bound for the discrete parameterization is equal to the variance of the GMM estimator if the discrete approximation is fine enough.

Much work has been done on these methods since these seminal contributions, and much remains in progress. This article discusses some of the developments since the work of Hansen (1982). In particular, it focuses on some of the recent work on empirical likelihood–type estimators, where Hansen again made important early contributions (Hansen, Heaton, and Yaron 1996, in a special issue of this journal on small-sample properties of GMM estimators), and which continues to be a area of great research activity. This literature developed partly in response to criticisms of the small-sample properties of the two-step GMM estimator. Researchers found in a number of studies that with the degree of overidentification

Journal of Business & Economic Statistics
October 2002, Vol. 20, No. 4
DOI 10.1198/073500102288618630

high, these estimators had substantial biases, and confidence intervals had poor coverage rates (see, among others, Altonji and Segal 1996; Burnside and Eichenbaum 1996; Pagan and Robertson 1997). These findings are related to the results in the instrumental variables literature that with many or weak instruments, two-stage least squares can have poor properties (e.g., Bekker 1994; Bound, Jaeger, and Baker 1995; Staiger and Stock 1997; Stock, Wright, and Yogo 2002). Simulations, as well as theoretical results, suggest that the new estimators have limited information maximum likelihood (LIML)–like properties and lead to improved large-sample properties, at the expense of some computational cost. Both the advantages and costs of these methods are discussed herein. Because the research on GMM approaches has expanded so much since 1982, not all areas can be discussed in great detail. In particular, time series settings are not addressed. Another area where a great deal of work has been done concerns the case with nonsmooth moment functions. Following the work on quantile regression by Koenker and Bassett (1978), much research has allowed for quantile-type moment functions, in just-identified or overidentified settings (e.g., Powell 1984; Honoré 1992). In addition, bootstrapping is not discussed. Here much work has been done (e.g., Brown and Newey 2002; Horowitz 2002), suggesting that the bootstrap can be an effective tool in improving coverage rates of confidence intervals in GMM settings.

The article is organized as follows. Section 2 emphasizes the richness of this framework by discussing some examples. Section 3 discusses the two-step estimators originally proposed by Hansen (1982). Then, Section 4 discusses alternative estimators for GMM settings based on empirical likelihood ideas that have recently attracted some attention, including the alternative proposals by Hansen et al. (1996). Section 5 explores higher-order differences between the various GMM estimators as a way of motivating choices between and them understanding their different large-sample properties. Section 6 discusses computational issues. Section 7 presents a small simulation study centered around an actual dataset that compares some GMM estimators. The choice of model is motivated partly by the higher-order comparisons that point to specific elements of the models that are relevant for the differences in finite samples. Section 8 concludes the article.

2. EXAMPLES

First, the generic form of the GMM estimation problem in a cross-sectional context is presented. The parameter vector θ^* is a K-dimensional vector, an element of Θ, which is a subset of $\mathbb{R}^K$. The random vector Z has dimension P, with its support $\mathcal{Z}$ a subset of $\mathbb{R}^P$. The moment function, $\psi : \mathcal{Z} \times \Theta \to \mathbb{R}^M$, is a known vector-valued function such that $E[\psi(Z, \theta^*)] = 0$ and $E[\psi(Z, \theta)] \neq 0$ for all $\theta \in \Theta$ with $\theta \neq \theta^*$. The researcher has available an iid random sample, $Z_1, Z_2, \ldots, Z_N$. Of interest are the properties of estimators for θ^* in large samples.

Many, if not most, models considered in econometrics fit this framework. Some examples are given next, but this list is by no means exhaustive.

2.1 Maximum Likelihood

If one specifies the conditional distribution of a variable Y given another variable X as $f_{Y|X}(y|x, \theta)$, then the score function satisfies these conditions for the moment function:

$$\psi(Y, X, \theta) = \frac{\partial \ln f}{\partial \theta}(Y|X, \theta).$$

By standard likelihood theory, the score function has expectation zero only at the true value of the parameter. Interpreting maximum likelihood estimators as GMM estimators suggests a way of deriving the covariance matrix under misspecification (e.g., White 1982), as well as an interpretation of the estimand in that case.

2.2 Instrumental Variables

Suppose that one has a linear model,

$$Y = X'\theta^* + \varepsilon,$$

with a vector of instruments, Z. In that case the moment function is

$$\psi(Y, X, Z, \theta) = Z \cdot (Y - X'\theta).$$

The validity of Z as an instrument, together with a rank condition, implies that θ^* is the unique solution to $E[\psi(Y, X, Z, \theta)] = 0$. This is a case in which the fact that the methods allow for more moments than unknown parameters is of great importance, because instruments are often independent of structural error terms, implying that any function of the basic instruments is orthogonal to the errors.

2.3 A Dynamic Panel Data Model

Consider the following panel data model with fixed effects:

$$Y_{it} = \eta_i + \theta Y_{it-1} + \varepsilon_{it},$$

where ε_{it} has mean 0 given $\{Y_{it-1}, Y_{it-2}, \ldots\}$. Observations Y_{it} are available for $t = 1, \ldots, T$ and $i = 1, \ldots, N$, with N large relative to T. This is a stylized version of the type of panel data models studied by Chamberlain (1992), Keane and Runkle (1992), and Blundell and Bond (1998). This specific model was previously studied by Bond, Bowsher, and Windmeijer (2001). Moment functions can be constructed by differencing and using lags as instruments, following Arellano and Bond (1991) and Ahn and Schmidt (1995):

$$\psi_{1t}(Y_{i1}, \ldots, Y_{iT}, \theta) = \begin{pmatrix} Y_{it-2} \\ Y_{it-3} \\ \vdots \\ Y_{i1} \end{pmatrix} \cdot ((Y_{it} - Y_{it-1} - \theta(Y_{it-1} - Y_{it-2})).$$

This leads to $t-2$ moment functions for each value of $t = 3, \ldots, T$, leading to a total of $(T-1) \cdot (T-2)/2$ moments with only a single parameter. One would typically expect the long lags to not necessarily contain much information, but they are often used to improve efficiency. In addition, under the assumption that the initial condition is drawn from the

stationary long-run distribution, the following additional $T-2$ moments are valid:

$$\psi_{2t}(Y_{i1}, \ldots, Y_{iT}, \theta) = (Y_{it-1} - Y_{it-2}) \cdot (Y_{it} - \theta Y_{it-1}).$$

Despite the different nature of the two sets of moment functions, which makes them potentially very useful in the case in which the autoregressive parameter is close to unity, they can be combined in the GMM framework.

2.4 Combining Data Sets

Suppose that one has an econometric model that fits in the GMM framework with moment function $\psi_1(Z, \theta)$. Now suppose that one also has direct information about the expectation or quantiles of functions of Z. This situation may arise if one combines data from a sample survey that comprises a random sample from some population and aggregate statistics from the same population. In this setting one can efficiently exploit this information by adding moment functions, as shown by Imbens and Lancaster (1994) and Hellerstein and Imbens (1999). For example, if the expectation of Z is known to be μ_Z, then one can add a moment function

$$\psi_2(Z) = Z - \mu_Z.$$

Note that in this case the second moment function does not depend on any unknown parameters. Nevertheless, through its correlation with the other moment functions, its presence can improve efficiency. A specific example is choice-based sampling (e.g., Manski and Lerman 1977; Cosslett 1981; Imbens 1992; Wooldridge 1999) or, more generally, stratified sampling, in which often it is assumed that population shares of choices or strata are known, as are possibly other characteristics of the population. This knowledge can be incorporated efficiently by adding moment functions describing this knowledge to the estimating equations that would used for estimation in the absence of such knowledge.

3. TWO-STEP GENERALIZED METHOD-OF-MOMENTS ESTIMATION

3.1 Estimation and Inference

In the just-identified case where M, the dimension of ψ, and K, the dimension of θ, are identical, one can generally estimate θ^* by solving

$$0 = \frac{1}{N}\sum_{i=1}^{N} \psi(Z_i, \hat{\theta}_{gmm}). \tag{1}$$

If the sample average is replaced by the expectation, then the unique solution is equal to θ^*, and under regularity conditions (e.g., Hansen 1982; Newey and McFadden 1994), solutions to (1) will be unique in large samples and consistent for θ^*. If $M > K$, the situation is more complicated, because there will be no solution to (1).

Hansen's (1982) solution was to generalize the optimization problem to the minimization of the quadratic form

$$Q_{C,N}(\theta) = \frac{1}{N}\left[\sum_{i=1}^{N} \psi(z_i, \theta)\right]' \cdot C \cdot \left[\sum_{i=1}^{N} \psi(z_i, \theta)\right], \tag{2}$$

for some positive definite $M \times M$ symmetric matrix C. Under the regularity conditions given by Hansen (1982) and Newey and McFadden (1994), the minimand $\hat{\theta}_{gmm}$ of (2) has the following large-sample properties:

$$\hat{\theta}_{gmm} \xrightarrow{p} \theta^*$$

and

$$\sqrt{N}(\hat{\theta}_{gmm} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma' C \Gamma)^{-1}\Gamma' C \Delta C \Gamma (\Gamma' C \Gamma)^{-1}).$$

In the just-identified case with the number of parameters K equal to the number of moments M, the choice of weight matrix C is immaterial, because $\hat{\theta}_{gmm}$ (at least in large samples) will be equal to the value of θ that sets the average moments exactly equal to 0. In that case Γ is a square matrix, and because it is full rank by assumption, Γ is invertible, and the asymptotic covariance matrix reduces to $(\Gamma'\Delta^{-1}\Gamma)^{-1}$ irrespective of the choice of C. In the overidentified case with $M > K$, however, the choice of the weight matrix C is important. In this case the optimal choice for C in terms of minimizing the asymptotic variance is the inverse of the covariance of the moments, Δ^{-1}. Using the optimal weight matrix, the asymptotic distribution is

$$\sqrt{N}(\hat{\theta}_{gmm} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma'\Delta^{-1}\Gamma)^{-1}). \tag{3}$$

This estimator is generally not feasible, because typically Δ^{-1} is not known to the researcher. The feasible solution proposed by Hansen (1982) is to obtain an initial consistent but generally inefficient estimate of θ^* by minimizing $Q_{C,N}(\theta)$ using an arbitrary positive definite $M \times M$ matrix C, for example, the identity matrix of dimension M. Given this initial estimate, $\tilde{\theta}$, one can estimate the optimal weight matrix as

$$\widehat{\Delta}^{-1} = \left[\frac{1}{N}\sum_{i=1}^{N} \psi(z_i, \tilde{\theta}) \cdot \psi(z_i, \tilde{\theta})'\right]^{-1}.$$

In the second step one estimates θ^* by minimizing $Q_{\widehat{\Delta}^{-1},N}(\theta)$. The resulting estimator, $\hat{\theta}_{gmm}$, has the same first-order asymptotic distribution as the minimand of the quadratic form with the true, rather than the estimated, optimal weight matrix, $Q_{\Delta^{-1},N}(\theta)$.

Hansen (1982) also suggested a specification test for this model. If the number of moments exceeds the number of free parameters, then not all average moments can be set equal to 0, and their deviation from 0 forms the basis of Hansen's test, similar to tests developed by Sargan (1958) (see also Newey 1985a, b). Formally, the test statistic is

$$T = \widehat{Q}_{\widehat{\Delta}^{-1},N}(\hat{\theta}_{gmm}).$$

Under the null hypothesis that all moments have expectation equal to 0 at the true value of the parameter, θ^*, the distribution of the test statistic converges to a chi-squared distribution with degrees of freedom equal to the number of overidentifying restrictions, $M-K$.

One can also interpret the two-step estimator for overidentified GMM models as a just-identified GMM estimator with an augmented parameter vector (e.g., Newey and McFadden

1994; Chamberlain and Imbens 2003). Define the following moment function:

$$h(x,\delta)=h(x,\theta,\Gamma,\Delta,\beta,\Lambda)=\begin{pmatrix}\Lambda-\frac{\partial\psi}{\partial\theta'}(x,\beta)\\ \Lambda' C\psi(x,\beta)\\ \Delta-\psi(x,\beta)\psi(x,\beta)'\\ \Gamma-\frac{\partial\psi}{\partial\theta'}(x,\theta)\\ \Gamma'\Delta^{-1}\psi(x,\theta)\end{pmatrix}. \quad (4)$$

Because the dimension of the moment function $h(\cdot)$, $M\times K+K+(M+1)\times M/2+M\times K+K=(M+1)\times(2K+M/2)$, is equal to the combined dimensions of its parameter arguments, the estimator for $\delta=(\theta,\Gamma,\Delta,\beta,\Lambda)$ obtained by setting the sample average of $h(\cdot)$ equal to 0 is a just-identified GMM estimator. The first two components of $h(x,\delta)$ depend only on β and Λ and have the same dimension as these parameters. Hence β^* and Λ^* are implicitly defined by the equations

$$E\left[\begin{pmatrix}\Lambda-\frac{\partial\psi}{\partial\theta'}(X,\beta)\\ \Lambda' C\psi(X,\beta)\end{pmatrix}\right]=0.$$

Given β^* and Λ^*, Δ^* is defined through the third component of $h(x,\delta)$, and given β^*, Λ^*, and Δ^*, the final parameters θ^* and Γ^* are defined through the last two moment functions.

This interpretation of the overidentified two-step GMM estimator as a just-identified GMM estimator in an augmented model is interesting, because it also emphasizes that results for just-identified GMM estimators, such as the validity of the bootstrap, can be directly translated into results for overidentified GMM estimators. In another example, using the standard approach to finding the large-sample covariance matrix for just-identified GMM estimators, one can use the just-identified representation to find the covariance matrix for the overidentified GMM estimator that is robust against misspecification: the appropriate submatrix of

$$\left(E\left[\frac{\partial h}{\partial\delta}(X,\delta^*)\right]\right)^{-1}E[h(Z,\delta^*)h(Z,\delta^*)']\left(E\left[\frac{\partial h}{\partial\delta}(Z,\delta^*)\right]\right)^{-1},$$

estimated by averaging at the estimated values. This is the GMM analog of the White (1982) covariance matrix for the maximum likelihood estimator under misspecification.

3.2 Efficiency

Chamberlain (1987) demonstrated that Hansen's (1982) estimator is efficient, not just in the class of estimators based on minimizing the quadratic form $Q_{C,N}(\theta)$, but also in the larger class of semiparametric estimators exploiting the full set of moment conditions. What is particularly interesting about this argument is the relation to the subsequent empirical likelihood literature. Many semiparametric efficiency bound arguments (e.g., Newey 1990; Hahn 1998) implicitly build fully parametric models that include the semiparametric one and then search for the least favorable parameterization. Chamberlain's argument is qualitatively different. He proposed a specific parametric model that can be made arbitrarily flexible, and thus arbitrarily close to the model that generated the data, but does not typically include that model. The advantage of Chamberlain's proposed model is that in some cases it is very convenient to work with, in the sense that its variance bound can be calculated in a straightforward manner. The specific model assumes that the data are discrete with finite support $\{\lambda_1,\ldots,\lambda_L\}$ and unknown probabilities $\pi_1,\ldots,\pi_L$. The parameters of interest are then implicitly defined as functions of these points of support and probabilities. With only the probabilities unknown, calculation of the variance bound on the parameters of the approximating model is conceptually straightforward. It then sufficies to translate that into a variance bound on the parameters of interest. If the original model is overidentified, then one has restrictions on the probabilities. These are again easy to evaluate in terms of their effect on the variance bound.

Given the discrete model, it is straightforward to obtain the variance bound for the probabilities, and thus for any function of them. The remarkable point is that one can rewrite these bounds in a way that does not involve the support points. This variance turns out to be identical to the variance of the two-step GMM estimator, thus proving its efficiency.

4. EMPIRICAL LIKELIHOOD

4.1 Background

A number of alternative estimators have been proposed in various forms and for various special cases. The motivation for these estimators and associated inference procedures is often the bias of the two-step GMM estimator, or the lack of accuracy of confidence intervals. For example, Altonji and Segal (1996) found large biases in a specific model with nonnormal moments. Pagan and Robertson (1997) listed a number of problems associated with finite-sample properties of GMM procedures, and discussed some potential solutions. Some of these alternative estimators have appealing information-theoretic interpretations in addition to being invariant to linear transformations of the moment functions. Specifically, Back and Brown (1990), Qin and Lawless (1994), Hansen et al. (1996), Imbens (1997), Smith (1997), and Kitamura and Stutzer (1997) have proposed estimators for the general overidentified GMM case that extends the empirical likelihood literature in statistics (Owen 1988; DiCiccio, Hall, and Romano 1991). (See Owen 2001 for a recent monograph on empirical likelihood.) A major advantage shared by all of these methods is that they have invariance properties. The two-step GMM estimator requires that the researcher make an initial choice about the weight matrix used in the first step. This choice affects the numerical values of the final estimates, even if this difference is of sufficiently low order that it does not affect the large-sample asymptotic distribution. Nevertheless, it introduces an ambiguity in the estimation procedure that can be avoided using the empirical likelihood methods.

To focus ideas, consider a random sample $Z_1,Z_2,\ldots,Z_N$ of size N from some unknown distribution. If one wishes to estimate the common distribution of these random variables, then the natural choice is the empirical distribution, which puts weight $1/N$ on each of the N sample points. However, this is not necessarily an appropriate estimate in a GMM setting. Suppose that the moment function is

$$\psi(z,\theta)=z,$$

implying that the expected value of Z is 0. Note that in this simple example, this moment function does not depend on any unknown parameter. The empirical distribution function with weights $1/N$ does not satisfy the restriction $E_F[Z] = 0$, because $E_{F_{emp}}[Z] = \sum z_i/N \neq 0$. The idea behind empirical likelihood is to modify the weights to ensure that the estimated distribution $\hat{F}$ does satisfy the restriction. In other words, the proper approach is to look for the distribution function closest to F_{emp} within the set of distribution functions satisfying $E_F[Z] = 0$. Empirical likelihood provides an operationalization of the concept of closeness here. The empirical likelihood is

$$\mathcal{L}(\pi_1, \ldots, \pi_N) = \prod_{i=1}^{N} \pi_i$$

for $0 \leq \pi_i \leq 1$, $\sum_{i=1}^{N} \pi_i = 1$. This is not a likelihood function in the standard sense and thus does not have all the properties of likelihood functions. The empirical likelihood estimator for the distribution function is

$$\max_{\pi} \sum_{i=1}^{N} \pi_i \qquad \text{subject to} \quad \sum_{i=1}^{N} \pi_i = 1 \quad \text{and} \quad \sum_{i=1}^{N} \pi_i z_i = 0.$$

Without the second restriction, the π's would be estimated to be $1/N$, but the second restriction forces them slightly away from $1/N$ in a way that ensures the restriction is satisfied. In this example, the solution for the Lagrange multiplier is the solution to the equation

$$\sum_{i=1}^{N} \frac{z_i}{1 + t \cdot z_i} = 0,$$

and the solution for π_i is

$$\hat{\pi}_i = 1/(1 + t \cdot z_i).$$

More generally, in the overidentified case a major focus is on obtaining point estimates through the following estimator for θ:

$$\max_{\theta, \pi} \sum_{i=1}^{N} \ln \pi, \quad \text{subject to} \quad \sum_{i=1}^{N} \pi_i = 1, \ \sum_{i=1}^{N} \pi_i \cdot \psi(z_i, \theta) = 0. \quad (5)$$

Qin and Lawless (1994) and Imbens (1997) showed that if the moment conditions are correctly specified, this estimator is equivalent, to order $O_p(N^{-1/2})$, to the two-step GMM estimator. This simple discussion illustrates that for some (and, in fact, many) purposes, the empirical likelihood function has the same properties as a parametric likelihood function. This idea, first proposed by Owen (1988), turns out to be very powerful with many applications. Owen (1988) showed how one can construct confidence intervals and hypothesis tests based on this notion. DiCiccio et al. (1991) showed that empirical likelihood is Bartlett correctable.

Related ideas have appeared in a number of places. Cosslett's (1981) work on choice-based sampling can be interpreted as maximizing a likelihood function that is the product of a parametric part coming from the specification of the conditional choice probabilities and an empirical likelihood function coming from the distribution of the covariates. (See Imbens 1992 for a connection between Cosslett's work and two-step GMM estimation.) As mentioned before, Chamberlain's (1987) efficiency proof essentially consists of calculating the distribution of the empirical likelihood estimator and showing its equivalence to the distribution of the two-step GMM estimator. (See Back and Brown 1990 and Kitamura and Stutzer 1997 for a discussion of the dependent case, and Mittelhammer et al. 2000 for a general discussion.)

In addition, two other estimators have been proposed. Altonji and Segal (1996) found in a setting with highly nonnormal moments that the correlation between the estimator for the weight matrix and the average moments leads to substantial bias. They found that splitting the sample and estimating the weight matrix on a separate part of the sample improves the bias considerably. Another estimator suggested by Hansen et al. (1996) consists of iterating between estimating the weight matrix and minimizing the quadratic form in the two-step GMM estimator. This is not to be confused with minimizing the quadratic form over the parameter in the average moments and the weight matrix simultaneously, as in the continuously updating estimator, which was also proposed by Hansen et al. (1996) and fits into the generalized empirical likelihood class, as shown by Newey and Smith (2001). The iterated GMM estimator does not appear to have the same higher-order bias properties as the continuously updating estimator although, like the continuously updating estimator but unlike the two-step estimator, it is invariant to starting values, provided that it converges.

4.2 Cressie–Read Discrepancy Statistics and Generalized Empirical Likelihood

In this section we consider a generalization of the empirical likelihood estimators based on modifications of the objective function. Corcoran (1998) (see also Imbens, Spady, and Johnson 1998), focused on the Cressie–Read discrepancy statistic, for fixed λ, as a function of two vectors p and q of dimension N (Cressie and Read 1984):

$$I_\lambda(p, q) = \frac{1}{\lambda \cdot (1 + \lambda)} \sum_{i=1}^{N} p_i \left[\left(\frac{p_i}{q_i} \right)^{\lambda} - 1 \right].$$

The Cressie–Read minimum discrepancy estimators are based on minimizing this difference between the empirical distribution, that is, the N-dimensional vector with all elements equal to $1/N$, and the estimated probabilities, subject to all the restrictions being satisfied.

$$\min_{\pi, \theta} I_\lambda(\iota/N, \pi), \quad \text{subject to} \quad \sum_{i=1}^{N} \pi_i = 1 \quad \text{and} \quad \sum_{i=1}^{N} \pi_i \cdot \psi(z_i, \theta) = 0.$$

If there are no binding restrictions, because the dimension of $\psi(\cdot)$ and θ agree (the just-identified case), then the solution for π is the empirical distribution it self, and $\pi_i = 1/N$. More generally, if there are overidentifying restrictions, then there is no solution for θ to $\sum_i \psi(z_i, \theta)/N = 0$, and so the solution for π_i is as close as possible to $1/N$ in a way that ensures an

exact solution to $\sum_i \pi_i \psi(z_i, \theta) = 0$. The precise way in which the notion "as close as possible" is implemented is reflected in the choice of metric through λ.

Three special cases of this class have received the most attention. One case is the empirical likelihood estimator itself, which can be interpreted as the case with $\lambda \to 0$. This has the nice interpretation that it is the exact maximum likelihood estimator if Z has a discrete distribution. It does not rely on the discreteness for its general properties, but this interpretation does suggest that it may have attractive large-sample properties.

The second case is the exponential tilting estimator with $\lambda \to -1$ (Imbens et al. 1998), whose objective function is equal to the empirical likelihood objective function with the roles of π and ι/N reversed. It can also be written as

$$\min_{\pi,\theta} \sum_{i=1}^{N} \pi_i \ln \pi_i, \quad \text{subject to} \quad \sum_{i=1}^{N} \pi_i = 1 \quad \text{and} \quad \sum_{i=1}^{N} \pi_i \psi(z_i, \theta) = 0.$$

The third case is the case with $\lambda = -2$, which was originally proposed by Hansen et al. (1996) as the solution to

$$\min_{\theta} \frac{1}{N}\left[\sum_{i=1}^{N} \psi(z_i, \theta)\right]' \cdot \left[\frac{1}{N}\sum_{i=1}^{N} \psi(z_i, \theta)\psi(z_i, \theta)'\right]^{-1} \cdot \left[\sum_{i=1}^{N} \psi(z_i, \theta)\right],$$

where the GMM objective function is minimized over the θ in the weight matrix as well as the θ in the average moments. Hansen et al. (1996) called this the "continuously updating estimator." Newey and Smith (2001) pointed out that this estimator fits in the Cressie–Read class.

Smith (1997) considered a more general class of estimators, which he called "generalized empirical likelihood (GEL) estimators," starting from a different perspective. For a given function $g(\cdot)$, normalized so that it satisfied $g(0) = 1$ and $g'(0) = 1$, consider the saddlepoint problem

$$\max_{\theta} \min_{t} \sum_{i=1}^{N} g(t'\psi(z_i, \theta)).$$

This representation is more attractive from a computational perspective, because it reduces the dimension of the optimization problem to $M + K$ rather than to a constrained optimization problem of dimension $K + N$ with $M + 1$ restrictions. There is a direct link between the t parameter in the GEL representation and the Lagrange multipliers in the Cressie–Read representation. Newey and Smith (2001) discussed how to choose $g(\cdot)$ for a given λ so that the corresponding GEL and Cressie–Read estimators agree.

In general, the differences between the estimators within this class are relatively small compared to the differences between them and the two-step GMM estimators. In practice, the choice between them is driven largely by computational issues, which are discussed in more detail in Section 5. The empirical likelihood estimator does have the advantage of its exact likelihood interpretation and the resulting optimality properties for its bias-corrected version (Newey and Smith 2001). On the other hand, Imbens et al. (1998) argued in favor of the exponential tilting estimator as its influence function stays bounded where as denominator in the probabilities in the empirical likelihood estimator can get large. In simulations researcher have encountered more convergence problems with the continuously updating estimator (e.g., Hansen et al. 1996; Imbens et al. 1998).

4.3 Testing

Associated with the empirical likelihood estimators are three tests for overidentiyfing restrictions that are similar to the classical trinity of the likelihood ratio, these Wald, and Lagrange multiplier tests. Here we briefly review the implementation of the three tests in the empirical likelihood context. The leading terms of all three tests are identical to the leading term of the test developed by Hansen (1982) based on the quadratic form in the average moments.

The first test is based on the value of the empirical likelihood function. The test statistic compares the value of the empirical likelihood function at the restricted estimates, the $\hat{\pi}_i$ with that at the unrestricted values, $\pi_i = 1/N$,

$$LR = 2 \cdot (L(\iota/N) - L(\hat{\pi})),$$

where

$$L(\pi) = \sum_{i=1}^{N} \ln \pi_i.$$

As in the parametric case, the difference between the restricted and unrestricted likelihood functions is multiplied by 2 to obtain, under regularity conditions (e.g., Newey and Smith 2001), a chi-squared distribution with degrees of freedom equal to the number of overidentifying restrictions for the test statistic under the null hypothesis.

The second test, similar to the Wald test, is based on the difference between the average moments and their probability limit under the null hypothesis zero. As in the standard GMM test for overidentifying restrictions (Hansen 1982), the average moments are weighted by the inverse of their covariance matrix,

$$Wald = Q_{\widehat{\Delta}^{-1}, N}(\hat{\theta}) = \frac{1}{N}\left[\sum_{i=1}^{N} \psi(z_i, \hat{\theta})\right]' \widehat{\Delta}^{-1} \left[\sum_{i=1}^{N} \psi(z_i, \hat{\theta})\right],$$

where $\widehat{\Delta}$ is an estimate of the covariance matrix

$$\Delta = E[\psi(Z, \theta^*)\psi(Z, \theta)'],$$

typically based on a sample average at some consistent estimator for θ^*,

$$\widehat{\Delta} = \frac{1}{N}\sum_{i=1}^{N} \psi(z_i, \hat{\theta})\psi(z_i, \hat{\theta})',$$

or sometimes a fully efficient estimator for the covariance matrix,

$$\widehat{\Delta} = \frac{1}{N}\sum_{i=1}^{N} \hat{\pi}_i \psi(z_i, \hat{\theta})\psi(z_i, \hat{\theta})'.$$

The standard GMM test uses an initial estimate of θ^* in the estimation of Δ, but with the empirical likelihood estimators, it is more natural to substitute the empirical likelihood estimator itself. The precise properties of the estimator for Δ do not affect the large-sample properties of the test, and, like the likelihood ratio test, the test statistic has in large samples a chi-squared distribution with degrees of freedom equal to the number of overidentifying restrictions.

The third test is based on the Lagrange multipliers t. In large samples, their variance is

$$V_t = \Delta^{-1} - \Delta^{-1}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}.$$

This matrix is singular, with rank equal to $M-K$. Thus one option is to compare the Lagrange multipliers to zero using a generalized inverse of their covariance matrix,

$$LM_1 = t'\left(\Delta^{-1} - \Delta^{-1}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}\right)^{-g} t.$$

This is not very attractive, because it requires the choice of a generalized inverse. An alternative is to use the inverse of Δ^{-1} itself, leading to the test statistic

$$LM_2 = t'\Delta t.$$

Because

$$\sqrt{N}\cdot t = V_t \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(z_i,\theta^*) + o_p(1)$$

and $V_t\Delta V_t = V_t V_t^{-g} V_t = V_t$, it follows that

$$LM_2 = LM_1 + o_p(1).$$

Imbens et al. (1998) found in their simulations that tests based on LM_2 perform better than those based on LM_1. In large samples, both have a chi-squared distribution with degrees of freedom equal to the number of overidentifying restrictions. Again this test can be used with any efficient estimator for t, and with the Lagrange multipliers based on any of the discrepancy measures.

Imbens et al. (1998) and Bond et al. (2001) investigate through simulations the small-sample properties of various of these tests. It appears that the Lagrange multiplier tests are often more attractive than the tests based on the average moments, although so far there is only limited evidence in specific models. One can use the same ideas for constructing confidence intervals that do not directly use the normal approximation to the sampling distribution of the estimator (see Smith 1997; Imbens and Spady 2002).

5. HIGHER-ORDER PROPERTIES OF GENERALIZED METHOD OF MOMENTS ESTIMATORS

The two-step GMM estimator and the various members of the Cressie–Read discrepancy statistic and GEL classes are all first-order efficient. To distinguish between them, it is necessary to calculate alternative approximations to their finite-sample distributions. These alternative asymptotics can be based on sequences involving adding moment conditions as the sample size increases, letting parameter values converge to limiting values as the sample size increases, or considering higher-order asymptotic approximations.

This section covers some recent results on higher-order approximations of the two-step and GEL estimators. Newey and Smith (2001) calculated the mean squared error for the general case up to order $1/N^2$. This reveals that the two-step GMM estimator has a number of additional terms that are absent in all GEL estimators (including the continuously updating GMM estimator). Some of these additional terms depend on the correlation between the derivatives and the moments, and increase in magnitude as the number of moments increase. In addition, they found that if the third moments of ψ are zero, then the various GEL estimators are the same up to order $1/N^2$.

Imbens and Spady (2001) considered the properties of bias and mean squared error of various estimators as the number of moments increases. These results are by necessity limited to the specific sequence chosen, but may offer some insight in the properties of the various estimators, because relatively few parameters are involved. Consider a sequence of independent and identically distributed pairs of random vectors $\{(v_i, w_i)\}_{i=1}^N$. The dimension of v_i and w_i is $M \geq 1$. Of interest is a scalar parameter, θ, satisfying

$$E[\psi(v_i, w_i, \theta)] = 0$$

for $i = 1, \ldots, N$, where

$$\psi(v_i, w_i, \theta) = (v_i + e_1)\cdot\theta - w_i = \begin{pmatrix} (v_{i1}+1)\cdot\theta - w_{i1} \\ v_{i2}\cdot\theta - w_{i2} \\ \vdots \\ v_{iM}\cdot\theta - w_{iM} \end{pmatrix}$$

and e_1 is an M-vector with the first element equal to 1 and the other elements equal to 0. To explore the properties of various estimators for θ as the degree of overidentification, $(M-1)$, increases, following Donald and Newey (2001), who examined the behavior of various instrumental variables estimators as the number of instruments increases, and Newey and Smith (2001), who looked at bias of GEL and GMM estimators, consider the leading terms in the asymptotic expansion of the estimators and the rate at which the moments of these terms increase with M.

Make the following simplifying assumptions. The pairs (v_{im}, w_{im}) and (v_{jn}, w_{jn}) are independent if either $i \neq j$ or $n \neq m$ (or both) and have the same distribution. Let $\mu_{rp} = E[v_{im}^r \cdot w_{im}^p]$ denote the moments of this distribution. Moments up to order $p+r \leq 6$ are assumed to be finite. Without essential loss of generality, let $\mu_{10} = \mu_{01} = 0$, implying that the true value of θ is $\theta^* = 0$, let $\mu_{20} = \mu_{02} = 1$, and let $\mu_{11} = \rho$ be the correlation coefficient of v_{im} and w_{im}. This is clearly a limited setup, although it may represent the essence of some of the poor small-sample behavior of some of the GMM estimators. Note that the fact that all of the first-order information about θ is in the first moment is without loss of generality. One can always reorganize the moments to ensure that only the derivative of the first moment differs from 0 in expectation, and that the other moments are uncorrelated with the first moment. Specifically, given a general moment function

$\psi(z, \theta)$ with the derivative matrix of the first K moments nonsingular in expectation, the moment vector can be rewritten as

$$\rho(z, \theta) = \begin{pmatrix} (\Gamma\Delta^{-1}\Gamma)^{-1/2}\Gamma'\Delta^{-1} \\ (\Delta_{22} - \Gamma_2(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma_2)^{-1/2}\begin{pmatrix}0 & I_{M-1}\end{pmatrix} - \Gamma_2(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1} \end{pmatrix} \times \psi(z, \theta),$$

where

$$\Delta = \begin{pmatrix}\Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22}\end{pmatrix} \quad \text{and} \quad \Gamma = \begin{pmatrix}\Gamma_1 \\ \Gamma_2\end{pmatrix},$$

are partitioned into the first K moments and the last $M - K$ moments. This rearranging makes the expected derivatives of moments other than the first K equal to 0, and ensures that all moments are uncorrelated,

$$E\left[\frac{\partial\rho}{\partial\theta'}(Z, \theta^*)\right] = \begin{pmatrix}\mathcal{I}_K \\ 0\end{pmatrix} \quad \text{and} \quad E[\rho(Z, \theta^*)\cdot\rho(Z, \theta^*)'] = \mathcal{I}_M,$$

where $\mathcal{I}_M$ is the identity matrix of dimension K. Premultiplying the moment vector by a nonsingular matrix does not affect the numerical values of the empirical likelihood estimators. It does, however, affect the numerical values of the two-step GMM estimators, which are not invariant to linear transformations of the moments. Of course, the assumptions here are stronger than merely assuming that the first-order information is in the first moment; all of the moments are assumed to be independent, not merely uncorrelated. In addition, all moments are assumed to be linear in the parameters.

Given this setup, interest focuses on the bias and mean squared error of the various estimators; in particular, how these change as a function of the number of moments M. Consider three estimators. The first estimator uses only the first informative moment and ignores the other moments, redundant in the terminology of Breusch, Qian, Schmidt, and Wyhowski (1999). In practice, of course, this estimator is not feasible—the researcher does not know which moments are relevant. Note that although it is possible to use selection criteria that in large samples select the informative set of moments and discard the others, such selection procedures generally affect the higher-order statistical properties of the estimators (see, e.g., Andrews 1999; Hall and Peixe 2001; Hall and Inoue 2001). Formally, the optimal GMM estimator is

$$\hat{\theta}_{opt} = \bar{w}_1/(1 + \bar{v}_1).$$

The second estimator is the two-step GMM estimator, in which the weight matrix is estimated at the true value of the parameter,

$$\hat{\theta}_{gmm} = \text{argmin}_\theta \left(\sum_{i=1}^N \psi(v_i, w_i, \theta)\right)' \times \left(\sum_{i=1}^N \psi(v_i, w_i, \theta^*)\psi(v_i, w_i, \theta^*)'\right)\left(\sum_{i=1}^N \psi(v_i, w_i, \theta)\right).$$

The third estimator is the exponential tilting estimator,

$$\min_{\pi,\theta} \sum_{i=1}^N \pi_i \ln \pi_i, \quad \text{subject to} \quad \sum_{i=1}^N \pi_i = 1 \quad \text{and} \quad \sum_{i=1}^N \pi_i \psi(v_i, w_i, \theta) = 0.$$

Determining the bias requires the expansions of these estimators up to order $O(N^{-1})$. Using the results of Imbens and Spady (2001) leads to

$$\hat{\theta}_{opt} = \bar{w}_1 - \bar{w}_1\bar{v}_1 + o_p(N^{-1}),$$

$$\hat{\theta}_{gmm} = \bar{w}_1 - 2\bar{v}_1\bar{w}_1 + (\overline{ww}'_{11} - 1)\cdot\bar{w}_1 - e_1'(\overline{ww}' - \mathcal{I}_M)\bar{w} + \bar{v}'\bar{w} + o_p(N^{-1}),$$

and

$$\hat{\theta}_{et} = \bar{w}_1 + (\overline{ww}'_{11} - 1)\bar{w}_1 - e_1'(\overline{ww}' - \mathcal{I}_M)\bar{w} + \bar{w}'\bar{v} - 2\bar{w}_1\bar{v}_1 - \rho\bar{w}'\bar{w} + \rho\bar{w}_1^2 + o_p(N^{-1}).$$

Using these expansions to calculate the bias of the leading terms yields

$$\text{bias}_{opt} = -\rho/N + o(1/N),$$

$$\text{bias}_{gmm} = -\rho/N + \rho(M-1)/N + o(1/N),$$

and

$$\text{bias}_{et} = -\rho/N + o(1/N).$$

The key result is that the bias of the GMM estimator increases linearly with the number of moments. In contrast, the bias of the exponential tilting estimator is not affected by the number of irrelevant moments up to the order considered here. The GMM bias is linear in the number of overidentifying restrictions, with the coefficient equal to the correlation between the moments and their derivatives. This bias arises from the term $\bar{w}'\bar{v}$. This term is also present in the exponential tilting estimator, but the bias it induces is offset by the presence of an additional term, $-\rho\bar{w}'\bar{w}$. The same holds for the other members of the generalized empirical likelihood class.

Note that the correlation between the moments and the derivatives is not the only potential source of bias. Altonji and Segal (1996) considered an example with significant biases, although in their case there is no correlation between the moments and their derivatives. In their case the bias arises from the estimation of the weight matrix. In general, however, if the derivatives are stochastic and correlated with the moments, then the bias will be dominated by the term that increases linearly in the number of moments.

6. COMPUTATIONAL ISSUES

The two-step GMM estimator requires two minimizations over a K-dimensional space. The empirical likelihood estimator in its original likelihood form (5) requires maximization over a space of dimension K (for the parameter θ) plus N (for the N probabilities), subject to $M + 1$ restrictions (on the M moments and the adding-up restriction for the probabilities). This is in general a much more formidable computational

problem than two optimizations in a K-dimensional space. A number of approaches to simplifying this problem have been attempted. This section discusses three of them in the context of the exponential tilting estimator, although most of them directly carry over to other members of the Cressie–Read or GEL classes.

6.1 Solving the First-Order Conditions

The first approach discussed here focuses on the first-order conditions and then concentrates out the probabilities π. This reduces the problem to one of dimension, $K+M$, K for the parameters of interest and M for the Lagrange multipliers for the restrictions. This is clearly a huge improvement, because the dimension of the problem no longer increases with the sample size. Let μ and t be the Lagrange multipliers for the restrictions $\sum \pi_i = 1$ and $\sum \pi_i \psi(z_i, \theta) = 0$. The first-order conditions for the π's and θ and the Lagrange multipliers are

$$0 = \ln \pi_i - 1 - \mu + t'\psi(z_i, \theta),$$

$$0 = \sum_{i=1}^{N} \pi_i \frac{\partial \psi}{\partial \theta'}(z_i, \theta),$$

$$0 = \exp(\mu - 1) \sum_{i=1}^{N} \exp\bigl(t'\psi(z_i, \theta)\bigr),$$

and

$$0 = \exp(\mu - 1) \sum_{i=1}^{N} \psi(z_i, \theta) \cdot \exp\bigl(t'\psi(z_i, \theta)\bigr).$$

The solution for π is

$$\pi_i = \exp(\mu - 1 + t'\psi(z_i, \theta)).$$

To determine the Lagrange multipliers t and the parameter of interest θ, one only needs π_i up to a constant of proportionality, so

$$0 = \sum_{i=1}^{N} \psi(z_i, \theta) \exp(t'\psi(z_i, \theta)) \tag{6}$$

and

$$0 = \sum_{i=1}^{N} t' \frac{\partial \psi}{\partial \theta}(z_i, \theta) \exp(t'\psi(z_i, \theta)) \tag{7}$$

can be solved. Solving the system of (6) and (7) is not straightforward. Because the probability limit of the solution for t is 0, the derivative with respect to θ of both first-order conditions converges at 0. Hence the matrix of derivatives of the first-order conditions converges to a singular matrix. As a result, standard approaches to solving systems of equations can behave erratically, and this approach to calculating $\hat{\theta}$ has been found to have poor operating characteristics.

6.2 Penalty Functions

Imbens et al. (1998) characterized the solution for θ and t as

$$\max_{\theta, t} K(t, \theta) \qquad \text{subject to} \quad K_t(t, \theta) = 0, \tag{8}$$

where $K(t, \theta)$ is the empirical analog of the cumulant generating function:

$$K(t, \theta) = \ln\left[\frac{1}{N} \sum_{i=1}^{N} \exp(t'\psi(z_i, \theta))\right].$$

They suggested solving this optimization problem by maximizing the unconstrained objective function with a penalty term that consists of a quadratic form in the restriction,

$$\max_{\theta, t} K(t, \theta) - 0.5 \cdot A \cdot K_t(t, \theta)' W^{-1} K_t(t, \theta), \tag{9}$$

for some positive definite $M \times M$ matrix W and a positive constant A. The first-order conditions for this problem are

$$0 = K_\theta(t, \theta) - A \cdot K_{t\theta}(t, \theta) W^{-1} K_t(t, \theta)$$

and

$$0 = K_t(t, \theta) - A \cdot K_{tt}(t, \theta) W^{-1} K_t(t, \theta).$$

For A large enough, the solution to this unconstrained maximization problem is identical to the solution to the constrained maximization problem (8). This follows from the fact that the constraint is in fact the first-order condition for $K(t, \theta)$. Thus, in contrast to many penalty function approaches, one does not have to let the penalty term go to infinity to obtain the solution to the constrained optimization problem; one need only let the penalty term increase sufficiently to make the problem locally convex. Imbens et al. (1998) suggested choosing

$$W = K_{tt}(t, \theta) + K_t(t, \theta) K_t(t, \theta)'$$

for some initial values for t and θ as the weight matrix, and reported that estimates are generally not sensitive to the choices of t and θ.

6.3 Concentrating out the Lagrange Multipliers

Mittelhammer, Judge, and Schoenberg (2001) suggested concentrating out both probabilities and Lagrange multipliers and then maximizing over θ without any constraints. As shown earlier, concentrating out the probabilities π_i can be done analytically. Although in general it is not possible to solve for the Lagrange multipliers t analytically, other than in the continuously updating case, for given θ, it is easy to numerically solve for t. This involves solving, in the exponential tilting case,

$$\min_{t} \sum_{i=1}^{N} \exp(t'\psi(z_i, \theta)).$$

This function is strictly convex as a function of t, with the easy-to-calculate first and second derivatives equal to

$$\sum_{i=1}^{N} \psi(z_i, \theta) \exp(t'\psi(z_i, \theta))$$

and

$$\sum_{i=1}^{N} \psi(z_i, \theta)\psi(z_i, \theta)' \exp(t'\psi(z_i, \theta)).$$

Therefore, concentrating out the Lagrange multipliers is computationally fast using a Newton–Raphson algorithm. The

resulting function $t(\theta)$ has derivatives with respect to θ equal to

$$\frac{\partial t}{\partial \theta'}(\theta) = -\left(\frac{1}{N}\sum_{i=1}^{N}\psi(z_i,\theta)\psi(z_i,\theta)'\exp(t(\theta)'\psi(z_i,\theta)\right)^{-1} \cdot \left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \psi}{\partial \theta'}(z_i,\theta)\exp(t(\theta)'\psi(z_i,\theta) + \psi(z_i,\theta)t(\theta)'\frac{\partial \psi}{\partial \theta'}(z_i,\theta)\exp(t(\theta)'\psi(z_i,\theta))\right)$$

After solving for $t(\theta)$, one can solve

$$\max_{\theta}\sum_{i=1}^{N}\exp(t(\theta)'\psi(z_i,\theta)). \tag{10}$$

Mittelhammer et al. (2001) used methods that do not require first derivatives to solve (10). This is not essential. Calculating first derivatives of the concentrated objective function only requires first derivatives of the moment functions, both directly and indirectly through the derivatives of $t(\theta)$ with respect to θ. In general, these are straightforward to calculate and likely to improve the performance of the algorithm.

In this method, in the end the researcher only must solve one optimization in a K-dimensional space, with the provision that for each evaluation of the objective function, he or she needs to numerically evaluate the function $t(\theta)$ by solving a convex maximization problem. The latter is fast, especially in the exponential tilting case, so that although the resulting optimization problem is arguably still more difficult than the standard two-step GMM problem, in practice it is not much slower. The simulations that follow use this method for calculating the estimates. After concentrating out the Lagrange multipliers are concentrated out using a Newton–Rahpson algorithm that uses both first and second derivatives, a Davidon–Fletcher–Powell algorithm is used to maximize over θ, using analytic first derivatives. Given a direction, a line search algoritm based on repeated quadratic approximations is used.

7. A DYNAMIC PANEL DATA MODEL

This section compares some of the GMM methods in the context of a panel data model discussed briefly in Section 2,

$$Y_{it} = \eta_i + \theta Y_{it-1} + \varepsilon_{it},$$

where ε_{it} has mean 0 given $\{Y_{it-1}, Y_{it-2}, \dots\}$. These are observations Y_{it} for $t = 1, \dots, T$ and $i = 1, \dots, N$, with N large relative to T. This is a stylized version of the type of panel data model studied extensively in the literature. Bond et al. (2001) studied this and similar models to evaluate the performance of test statistics based on different GMM and GEL estimators. Using the moments

$$\psi_{1t}(Y_{i1},\dots,Y_{iT},\theta) = \begin{pmatrix} Y_{it-2} \\ Y_{it-3} \\ \vdots \\ Y_{i1} \end{pmatrix} \cdot ((Y_{it} - Y_{it-1} - \theta(Y_{it-1} - Y_{it-2}))$$

leads to $t-2$ moment functions for each value of $t = 3, \dots, T$, and then to a total of $(T-1)\cdot(T-2)/2$ moments. In addition, under the assumption that the initial condition is drawn from the stationary long-run distribution, the following additional $T-2$ moments are valid:

$$\psi_{2t}(Y_{i1},\dots,Y_{iT},\theta) = (Y_{it-1} - Y_{it-2})\cdot(Y_{it} - \theta Y_{it-1}).$$

It is important to note, given the results discussed in Section 4, that the derivatives of these moments are stochastic and potentially correlated with the moments themselves. As a result, there is potentially a substantial difference between the different estimators, especially when the degree of overidentification is high.

The data used are from Abowd and Card (1989), taken from the PSID (see also Card 1994). This dataset comprises earnings data for 1,434 individuals for 11 years. The individuals are selected on having positive earnings in each of the 11 years, and their earnings are modeled in logarithms. Table 1 gives summary statistics for the data. There is much persistence in these earnings data, with the correlation coefficient after 10 years still .44, down from around .80 after 1 year. A key question is how much of that is due to permanent components (captured in the fixed effect) versus transitory components (captured in the autoregressive coefficient θ).

First, the model is estimated using only data from year 1 to t, for $t = 3, 4, \dots, 11$. For each of these nine datasets, θ and its standard error are estimated using the two-step GMM estimators, the Exponential Tilting (ET) estimator, and the iterated GMM (IGMM) estimator. These results are reported in Table 2.

Next, artificial datasets are generated to investigate the repeated sampling properties of these estimators. Two questions are of most interest. First, how do the median bias and median absolute error deteriorate as a function of the degree of overidentification? Here, unlike in the theoretical discussion in Section 4, as the number of years in the panel are increased, the additional moments do contain information, so they may in fact increase precision. At the same time, however, based on the theoretical calculations, the accuracy of the asymptotic approximations for a fixed sample size would be expected to deteriorate with the number of years. The second

Table 1. Summary Statistics for the Abowd–Card Log Earnings Data

	Year										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Average	8.79	8.83	8.86	8.92	8.97	8.92	8.89	8.92	8.95	8.95	8.90
Standard deviation	.71	.64	.63	.62	.59	.62	.63	.68	.66	.64	.71
Correlations											
1	1.00	.82	.71	.68	.67	.61	.57	.53	.55	.50	.44
2		1.00	.80	.76	.74	.69	.65	.61	.62	.57	.51
3			1.00	.78	.75	.70	.66	.61	.62	.58	.53
4				1.00	.82	.76	.72	.66	.70	.65	.59
5					1.00	.82	.73	.69	.71	.67	.60
6						1.00	.79	.73	.73	.70	.62
7							1.00	.75	.73	.68	.62
8								1.00	.78	.71	.64
9									1.00	.79	.71
10										1.00	.83
11											1.00

NOTE: Summary statistics are based on a sample of 1,434 individuals. The basic data are log earnings.

Table 2. Estimation Results for the Abowd–Card Log Earnings Data

		Number of time periods								
		3	4	5	6	7	8	9	10	11
GMM	Estimate	.43	.33	.32	.29	.30	.29	.30	.29	.29
	(Standard error)	(.12)	(.05)	(.04)	(.03)	(.03)	(.03)	(.02)	(.02)	(.02)
IGMM	Estimate	.43	.33	.32	.28	.28	.27	.28	.25	.24
	(Standard error)	(.12)	(.05)	(.04)	(.03)	(.03)	(.03)	(.02)	(.02)	(.02)
ET	Estimate	.43	.33	.33	.28	.29	.27	.27	.25	.24
	(Standard error)	(.12)	(.05)	(.04)	(.03)	(.03)	(.03)	(.02)	(.02)	(.02)

NOTE: Estimates of autoregressive coefficient for three estimators and samples with different numbers of waves.

question of interest is the performance of the confidence intervals for the parameter of interest. In two-stage least squares settings, it was found that with many weak instruments, the performance of standard confidence intervals varied widely between LIML and two-stage least squares estimators. Given the analogy drawn by Hansen et al. (1996) between the continuously updating estimator and LIML, the question arises of how the confidence intervals differ between two-step GMM and the various Cressie–Read and GEL estimators.

Using the Abowd–Card data, θ and the variance of the fixed effect and the idiosyncratic error term are estimated. The latter two are estimated to be around .3. Then data-generating processes are considered where the fixed effect has mean 0 and standard deviation equal to .3, and the error term has mean 0 and standard deviation .3. Two values for θ, .5 and .9, are explored. The value .5 is close to the value estimated from the Abowd–Card data. The second value is chosen closer to unity to make the instruments weaker, and thus to potentially worsen the performance of all estimators. In the $\theta = .9$ case, inference would be expected to be less reliable. It is of interest to assess the relative performance of the estimators in both cases. Also compare are results for the case where all moments are included and the case where only the first set of $(T-1)\cdot(T-2)/2$ moments is included, because often researchers do not wish to rely on stationarity of the initial conditions. For each dataset, θ is estimated using the first t years of data, for $t = 3, \ldots, 11$ using both the two-step GMM and the ET estimators.

Tables 3–5 report for the four data-generating processes the results for the three estimators, for each of the nine different panel lengths, the median bias, the median absolute error, both in levels and relative to the asymptotic standard error, and also

Table 3. Simulations for the Two-Step GMM Estimator

	Number of time periods								
	3	4	5	6	7	8	9	10	11
ψ_1(●) only									
$\theta = .9$									
Median bias	.03	−.11	−.09	−.07	−.05	−.05	−.04	−.03	−.03
Relative median bias	.04	−.33	−.43	−.52	−.57	−.62	−.61	−.71	−.68
Median absolute error	.46	.21	.14	.10	.07	.06	.05	.04	.04
Coverage rate 90% CI	.92	.86	.85	.84	.81	.82	.79	.76	.78
Coverage rate 95% CI	.94	.91	.92	.90	.88	.88	.88	.86	.86
All moments									
$\theta = .9$									
Median bias	−.00	.00	.00	.00	.00	.00	.00	.00	.00
Relative median bias	−.02	.08	.03	.08	.03	.11	.08	.13	.11
Median absolute error	.04	.03	.02	.02	.02	.01	.01	.01	.01
Coverage rate 90% CI	.88	.85	.82	.80	.80	.79	.78	.79	.76
Coverage rate 95% CI	.92	.91	.89	.87	.85	.86	.86	.88	.84
ψ_1(●) only									
$\theta = .5$									
Median bias	.01	−.00	−.00	−.00	−.00	−.00	−.00	−.00	−.00
Relative median bias	.09	−.07	−.10	−.06	−.12	−.12	−.15	−.19	−.11
Median absolute error	.08	.04	.03	.02	.02	.02	.01	.01	.01
Coverage rate 90% CI	.93	.89	.91	.91	.90	.89	.87	.88	.88
Coverage rate 95% CI	.97	.94	.95	.95	.94	.94	.94	.93	.94
All moments									
$\theta = .5$									
Median bias	−.00	.00	.00	−.00	−.00	−.00	.00	.00	.00
Relative median bias	−.07	.01	.03	−.06	−.08	−.00	.09	.11	.14
Median absolute error	.05	.03	.02	.01	.01	.01	.01	.01	.01
Coverage rate 90% CI	.91	.88	.88	.91	.91	.90	.89	.90	.90
Coverage rate 95% CI	.95	.94	.94	.95	.96	.96	.95	.94	.94

NOTE: The relative median bias reports the median bias divided by the large-sample standard errors. All results are based on 10,000 replications.

Table 4. Simulations for the Iterated GMM Estimator

	Number of time periods								
	3	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
$\psi_1(\theta)$ only									
$\theta = .9$									
Median bias	−.00	−.11	−.09	−.07	−.05	−.04	−.04	−.03	−.03
Relative median bias	−.00	−.34	−.49	−.57	−.59	−.63	−.66	−.75	−.69
Median absolute error	.46	.21	.14	.10	.07	.06	.05	.04	.04
Coverage rate 90% CI	.91	.86	.85	.83	.82	.83	.79	.78	.79
Coverage rate 95% CI	.94	.90	.91	.90	.88	.89	.88	.87	.87
All moments									
$\theta = .9$									
Median bias	−.00	.00	.00	.00	.00	.00	.00	.00	.00
Relative median bias	−.04	.00	.02	.09	.05	.13	.11	.13	.14
Median absolute error	.04	.03	.02	.02	.02	.02	.01	.01	.01
Coverage rate 90% CI	.89	.84	.80	.79	.78	.77	.76	.76	.74
Coverage rate 95% CI	.93	.89	.87	.85	.83	.83	.83	.84	.83
$\psi_1(\theta)$ only									
$\theta = .5$									
Median bias	.01	−.00	−.01	−.00	−.00	−.00	−.00	−.00	−.00
Relative median bias	.06	−.07	−.12	−.07	−.12	−.11	−.15	−.18	−.12
Median absolute error	.08	.04	.03	.02	.02	.01	.01	.01	.01
Coverage rate 90% CI	.93	.90	.91	.90	.91	.89	.88	.88	.88
Coverage rate 95% CI	.96	.95	.95	.95	.95	.95	.94	.93	.94
All moments									
$\theta = 0.5$									
Median bias	−.00	.00	.00	−.00	.00	.00	.00	.00	.00
Relative median bias	−.03	.01	.03	−.01	.01	.10	.12	.13	.14
Median absolute error	.05	.02	.02	.01	.01	.01	.01	.01	.01
Coverage rate 90% CI	.91	.89	.89	.90	.90	.89	.88	.88	.88
Coverage rate 95% CI	.95	.94	.94	.95	.94	.94	.94	.94	.94

NOTE: The relative median bias reports the bias divided by the large-sample standard error. All results are based on 10,000 replications.

Table 5. Simulations for Exponential Tilting Estimator

	Number of time periods								
	3	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
$\psi_1(\theta)$ only									
$\theta = .9$									
Median bias	.03	−.01	−.00	.00	.00	.00	0.01	.00	.00
Relative median bias	.04	−.03	−.01	.04	.04	.05	.11	.02	.08
Median absolute error	.46	.24	.14	.10	.07	.05	.05	.04	.03
Coverage rate 90% CI	.92	.91	.92	.89	.88	.88	.85	.88	.85
Coverage rate 95% CI	.94	.94	.96	.95	.95	.93	.93	.93	.92
All moments									
$\theta = .9$									
Median bias	.00	.00	.00	−.00	.00	.00	−.00	.00	.00
Relative median bias	.04	.09	.02	−.00	.01	.01	−.02	.08	.13
Median absolute error	.05	.03	.03	.02	.02	.01	.01	.01	.01
Coverage rate 90% CI	.87	.86	.84	.86	.88	.86	.87	.88	.87
Coverage rate 95% CI	.91	.90	.90	.91	.93	.92	.91	.93	.93
$\psi_1(\theta)$ only									
$\theta = .5$									
Median bias	.01	−.00	−.00	.00	.00	.00	.00	−.00	.00
Relative median bias	.09	−.02	−.04	.04	.00	.02	.02	−.03	.08
Median absolute error	.08	.05	.03	.02	.02	.02	.01	.01	.01
Coverage rate 90% CI	.93	.89	.91	.91	.90	.88	.87	.88	.90
Coverage rate 95% CI	.97	.94	.95	.95	.95	.95	.94	.94	.94
All moments									
$\theta = .5$									
Median bias	−.00	−.00	−.00	−.00	−.00	−.00	.00	.00	.00
Relative median bias	−.04	−.02	−.01	−.09	−.07	−.01	.02	.12	.10
Median absolute error	.05	.03	.02	.01	.01	.01	.01	.01	.01
Coverage rate 90% CI	.90	.87	.89	.90	.92	.91	.90	.90	.91
Coverage rate 95% CI	.95	.94	.94	.96	.95	.96	.95	.94	.95

NOTE: The relative median bias reports the bias divided by the large sample standard error. All results are based on 10,000 replications.

the coverage rate of the 90% and 95% confidence intervals. Table 3 presents the results for the two-step GMM estimator; Table 4, results for the iterated GMM estimator, and Table 5, results for the exponential tilting estimator. With the high autoregressive coefficient, $\theta = .9$, and using only the moments based on lagged values as instruments for the difference, the two-step and iterated GMM estimators have substantial bias and poor coverage rates. Relative to the asymptotic standard error, the bias is on the order of 50%–60%. With the larger set of moments, the bias goes down, but the coverage rate still is poor. With the lower value for the autoregressive coefficient, the bias and coverage rate are much better. The exponential tilting estimator does much better with the high autoregressive coefficient. The bias is small, on the order of 10% of the standard error, and the coverage rate is much closer to the nominal rate. This does not change if the autoregressive coefficient is .5, or if the larger set of moments is used. In this setting, the exponential tilting estimator is clearly superior.

8. CONCLUSION

Much work has been done since Hansen's (1982) seminal article formalizing a set of methods now known as GMM. This has become a crucial organizing principle for much of point estimation and inference in modern econometrics and is now a important part of any current graduate textbook or education. This article has attempted to describe some recent developments in this area that build on Hansen's work, and in particular some of the recent empirical likelihood estimators. The estimators use the same set of moments but remove some of the ambiguity stemming from the weight matrix estimation. In doing so, they address some of the problems associated with GMM estimation. These and other developments demonstrate that 20 years after Hansen's original contribution, work on GMM remains is a vibrant area of research. It is likely to remain so for the next 20 years.

ACKNOWLEDGMENTS

The author thanks Whitney Newey and Richard Spady for discussions during the preparation of this article, and the editors for comments on an earlier version.

[Received April 2002. Revised June 2002.]

REFERENCES

Abowd, J., and Card, D. (1989), "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57, 441–445.

Ahn, S., and Schmidt, P. (1995), "Efficient Estimation of Models for Dynamic Panel Data," *Journal of Econometrics*, 68, 5–28.

Altonji, J., and Segal, L. (1996), "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics,* 14, 353–366.

Andrews, D. (1999), "Consistent Moment Selection for Generalized Method of Moments Estimation," *Econometrica*, 67, 543–564.

Arellano, M., and Bond, S. (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economics and Statistics*, 58, 277–298.

——— (1993), "Implied Probabilities in GMM Estimators," *Econometrica*, 61, 971–976.

Bekker, P. (1994), "Alternative Approximations to the Distributions of Instrumental Variables Estimators," *Econometrica*, 62, 657–681.

Blundell, R., and Bond, S. (1998), "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87, 115–143.

Bond, S., Bowsher, C., and Windmeijer, F. (2001), "Criterion-Based Inference for GMM in Linear Dynamic Panel Data Models," IFS, London.

Bound, J., Jaeger, D., and Baker, R. (1995), "On Potential Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Breusch, T., Qian, H., Schmidt, P., and Wyhowski, D. (1999), "Redundancy of Moment Conditions," *Journal of Econometrics*, 91, 89–111.

Brown, B., and Newey, W., (2002), "Generalized Method of Moments, Efficient Bootstrapping, and Improved Inference," *Journal of Business and Economic Statistics*, 20, 507–517.

Burnside, C., and Eichenbaum, M. (1996), "Small-Sample Properties of Generalized Method of Moments-Based Wald Tests," *Journal of Business and Economic Statistics*, 14, 294–308.

Card, D. (1994), "Intertemporal Labour Supply: An Assessment," in *Advances in Econometrics*, ed. C. A. Sims, Cambridge, UK: Cambridge University Press.

Chamberlain, G. (1987), "Asymptotic Efficiency in Estimation With Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

——— (1992), "Sequential Moment Restrictions in Panel Data—Comment," *Journal of Business and Economic Statistics*, 10, 20–26.

Chamberlain, G., and Imbens, G. (2003), "Nonparametric Applications of Bayesian Inference," *Journal of Business and Economic Statistics*,

Corcoran, S. (1998), "Bartlett Adjustment of Empirical Discrepancy Statistics," *Biometrika*, 85, 967–972.

Cosslett, S. R. (1981), "Maximum Likelihood Estimation for Choice-Based Samples," *Econometrica*, 49, 1289–1316,

Cressie, N., and Read, T. (1984), "Multinomial Goodness-of-Fit Tests," *Journal of the Royal Statistical Society*, Ser. B, 46, 440–464.

Davidson, R., and MacKinnon, J. G. (1993), *Estimation and Inference in Econometrics*, New York: Oxford University Press.

Dicicio, T., Hall, P., and Romano, J. (1991), "Empirical Likelihood is Bartlett-Correctable," *The Annals of Statistics*, 19, 1053–1061.

Donald, S., and Newey, W. (2001), "Choosing the Number of Instruments," *Econometrica*, 69, 1161–1191.

Gallant, R. (1987), *Nonlinear Statistical Models*, New York: Wiley.

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

Hall, A., and Inuoue, A. (2001), "A Canonical Correlations Interpretation of Generalized Method of Moments Estimation With Applications to Moment Selection," unpublished manuscript, North Carolina State University, Dept. of Economics.

Hall, A., and Peixe, F. (2001), "A Consistent Method for the Selection of Relevant Instruments" unpublished manuscript, North Carolina State University, Dept. of Economics.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton, NJ: Princeton University Press.

Hansen, L.-P. (1982), "Large-Sample Properties of Generalized Method of Moment Estimators," *Econometrica*, 50, 1029–1054.

Hansen, L.-P., Heaton, J., and Yaron, A. (1996), "Finite Sample Properties of Some Alternative GMM Estimators," *Journal of Business & Economic Statistics*, 14, 3, 262–280.

Hayashi, F. (1999), *Econometrics*, Princeton, NJ: Princeton University Press.

Hellerstein, J., and Imbens, G. (1999), "Imposing Moment Restrictions From Auxiliary Data by Weighting," *Review of Economics and Statistics*, 81, 1–14.

Honoré, B. (1992), "Trimmed LAD and Least-Squares Estimation of Truncated and Censored Regression Model With Fixed Effects," *Econometrica*, 60, 533–565.

Horowitz, J. (2002), "The Bootstrap in Econometrics," in *Handbook of Econometrics*, Vol. 5, eds. J. Heckman and E. Leamer, Amsterdam: North-Holland.

Imbens, G. W. (1992), "An Efficient Method of Moments Estimator for Discrete Choice Models With Choice-Based Sampling," *Econometrica*, 60, 1187–1214.

Imbens, G. (1997). "One-step Estimators for Over-identified Generalized Method of Moments Models," *Review of Economic Studies* 64, 359–383.

Imbens, G. W., Spady, R. H., and Johnson, P. (1998), "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333–357.

Imbens, G. W., and Lancaster, T. (1994), "Combining Micro and Macro Data in Microeconometric Models," *Review of Economic Studies*, 61, 655–680.

Imbens, G., and Spady, R. (2001), "The Performance of Empirical Likelihood and Its Generalizations," unpublished working paper, University of California, Berkeley, Department of Economics.

——— (2002), "Confidence Intervals in Generalized Method of Moments Models," *Journal of Econometrics*, 107, 87–98.

Keane, M., and Runkle, D. (1992), "On the Estimation of Panel Data Models With Serial Correlation When Instruments Are not Strictly Exogenous," *Journal of Business and Economic Statistics*, 10, 1–9.

Kitamura, Y., and Stutzer, M. (1997), "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65, 861–874.

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.

Manski, C., and Lerman, S. (1977), "The Estimation of Choice Probabilities from Choice-Based Samples," *Econometrica*, 45, 1977–1988.

Mittelhammer, R., Judge, G., and Miller, D. (2000), *Econometric Foundations*, Cambridge, UK: Cambridge University Press.

Mittelhammer, R., Judge, G., and Schoenberg, R. (2001), "Empirical Evidence Concerning the Finite Sample Performance of EL-Type Structural Equation Estimation and Inference Methods," unpublished manuscript, University of California Berkeley.

Newey, W. (1985a), "Maximum Likelihood Specification Testing and Conditional Moment Tests," *Econometrica*, 53, 1047–1069.

——— (1985b), "Generalized Method of Moments Specification Testing," *Journal of Econometrics*, 29, 229–256.

——— (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

Newey, W., and McFadden, D. (1994), "Estimation in Large Samples," in *The Handbook of Econometrics*, Vol. 4, eds. D. McFadden and R. F. Engle, Amsterdam: North Holland.

Newey, W., and Smith, R. (2001), "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," mimeo, Massachusetts Institute of Technology.

Owen, A. (1988), "Empirical Likelihood Ratios Confidence Intervals for a Single Functional," *Biometrika*, 75, 237–249.

——— (2001), *Empirical Likelihood*, London: Chapman and Hall.

Pagan, A., and Robertson, J. (1997), "GMM and Its Problems," unpublished manuscript, Australian National University.

Powell, J. (1984), "Least Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics*, 25, 303–325.

Qin, J., and Lawless, J. (1994), "Generalized Estimating Equations," *The Annals of Statistics*, 20, 300–325.

Ruud, P. (2000), *An Introduction to Classical Econometric Theory*, Oxford University Press, Oxford.

Sargan, J. D. (1958), "The Estimation of Economic Relationships Using Instrumental Variables," *Econometrica*, 26, 393–415.

Smith, R. (1997), "Alternative Semiparametric Likelihood Approaches to Generalized Method of Moments Estimation," *Economic Journal*, 107, 503–519.

Staiger, D., and Stock, J. (1997), "Instrumental Variables Regression With Weak Instruments," *Econometrica*, 65, 557–586.

Stock, J., Wright, J., and Yogo, M. (2002), "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20, 518–529.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

Wooldridge, J. (1999), "Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples," *Econometrica*, 67, 1385–1406.

Wooldridge, J. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.